Introduction to Robotics for cognitive science

Dr. Andrej Lúčny

KAI FMFI UK

lucny@fmph.uniba.sk

Web page of the subject

www.agentspace.org/kv

Why is DL possible today and was not possible before?

Software inventions:

- Dropout & Batch normalization (solves overfitting & vanishing gradient)
- Xavier initialization
- Novel loss functions (metric loss function)

Hardware inventions:

- Big Data Storages
- Graphics Processing Units

Why is DL possible today and was not possible before?

Software inventions:

- Dropout & Batch normalization (solves overfitting & vanishing gradient)
- Xavier initialization
- Novel loss functions (metric loss function)

Hardware inventions:

- Big Data Storages
- Graphics Processing Units

We need powerful hardware not available everywhere, not only for training but also for the use of the DL models.

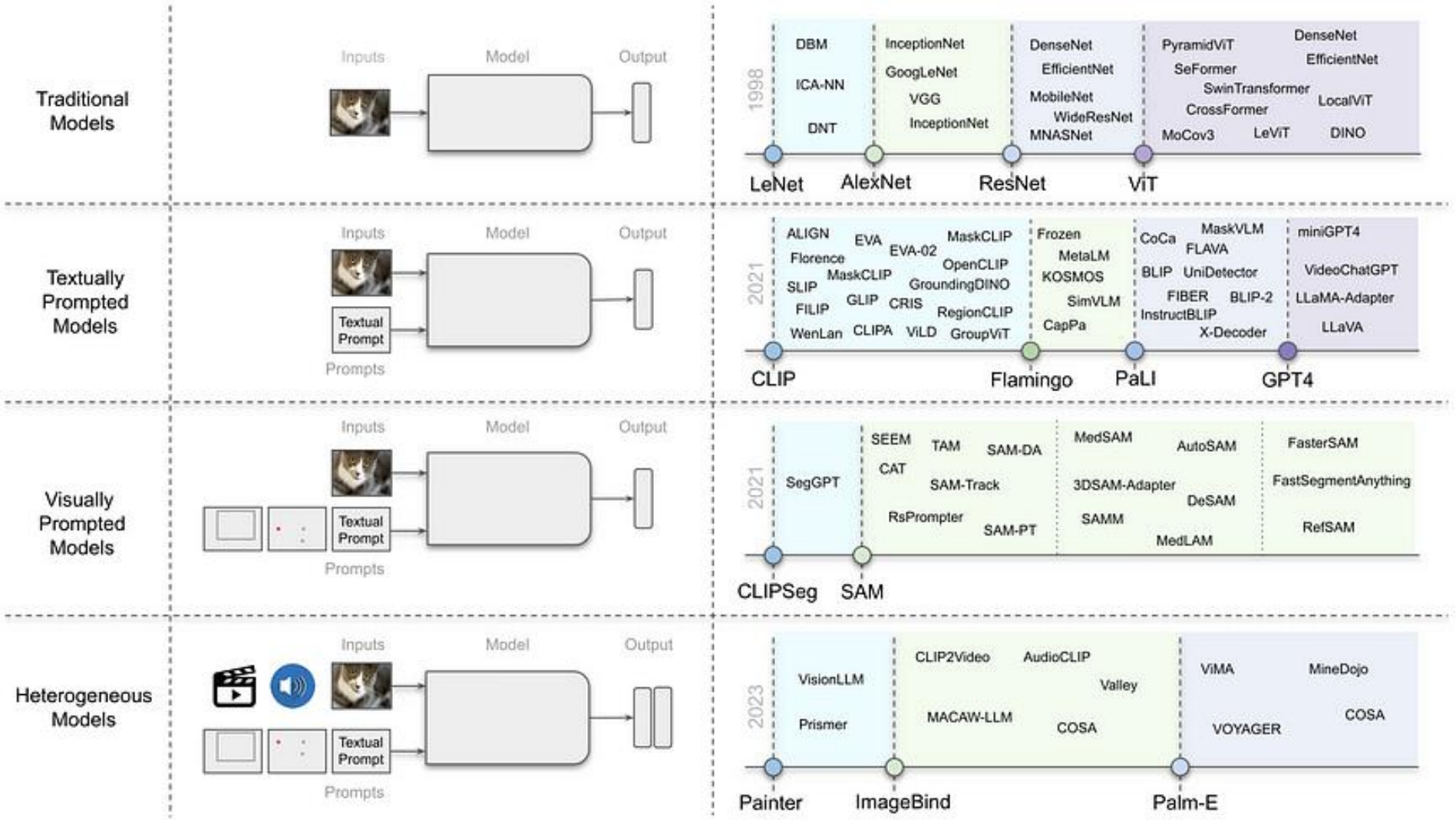
Foundation models

- Classify Anything
- Detect Anything
- Segment Anything
- Answer Anything

Know-how of humankind is stored in texts and images available on the Internet. The issue is to organize it and be able of employment – that is a job for deep learning

self-supervised + supervised

Foundation models



Solution: Cloud technology

- Instead of calling a local subroutine, program compose http request with attached marshaled arguments and get a marshaled result as a response

computer network cloud

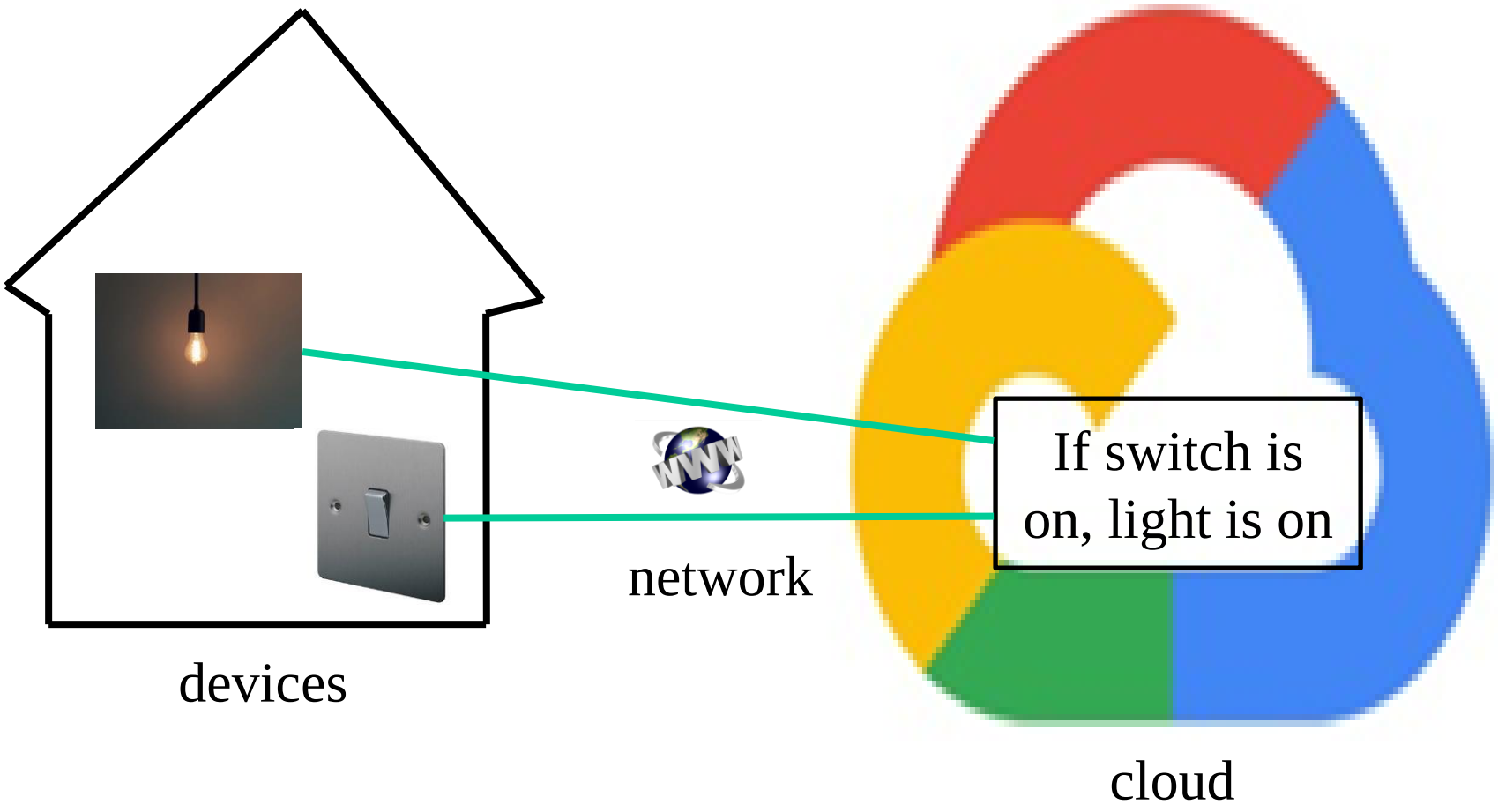
- Today such call takes 80 ms from EU, 40 ms from USA

Calling cognitive web services

```
api_key = "…"
with open("input.jpg", "rb") as image_file:
    base64_image = base64.b64encode(image_file.read()).decode('utf-8')
headers = {
    "Content-Type": "application/json",
    "Authorization": f"Bearer {api_key}"
}
payload = {
    "model": "gpt-4-vision-preview",
    "messages": [ {
        "role": "user",
        "content": [
            { "type": "text", "text": "Describe the image" },
            { "type": "image_url", "image_url": {
              "url": f"{base64_image}"
            }}
        ]
    } ],
    "max_tokens": 300
}
response = requests.post(
    "https://api.openai.com/v1/chat/completions",
    headers=headers, json=payload
)
print(response.json())
```

```
{'choices': [{'finish_reason': 'stop',
              'index': 0,
              'logprobs': None,
              'message': {'content': 'The image shows a man who appears to be '
                                     'giving a presentation or speech. He is '
                                     'dressed in business attire, featuring a '
                                     'dark suit, a light shirt, and a striped '
                                     'tie. The man is wearing glasses and has '
                                     'short hair. He is holding what appears '
                                     'to be a remote or a clicker for '
                                     'advancing slides in his right hand, '
                                     'suggesting he might be using visual aids '
                                     'or slides as part of his presentation. '
                                     'The background is somewhat blurred but '
                                     'seems to be a conference room or space '
                                     'suitable for lectures or corporate '
                                     "meetings. The man's expression and "
                                     'open-hand gesture indicate that he is '
                                     'engaged in communication, likely '
                                     'explaining or emphasizing a point.',
                          'role': 'assistant'}}],
 'created': 1715601422,
 'id': 'chatcmpl-9OOacV7HNkpWzgEV0LbIauE7lbwyx',
 'model': 'gpt-4-1106-vision-preview',
 'object': 'chat.completion',
 'system_fingerprint': None,
 'usage': {'completion_tokens': 126,
           'prompt_tokens': 1115,
           'total_tokens': 1241}}
```

Internet of Things (IoT)



Robot Pepper

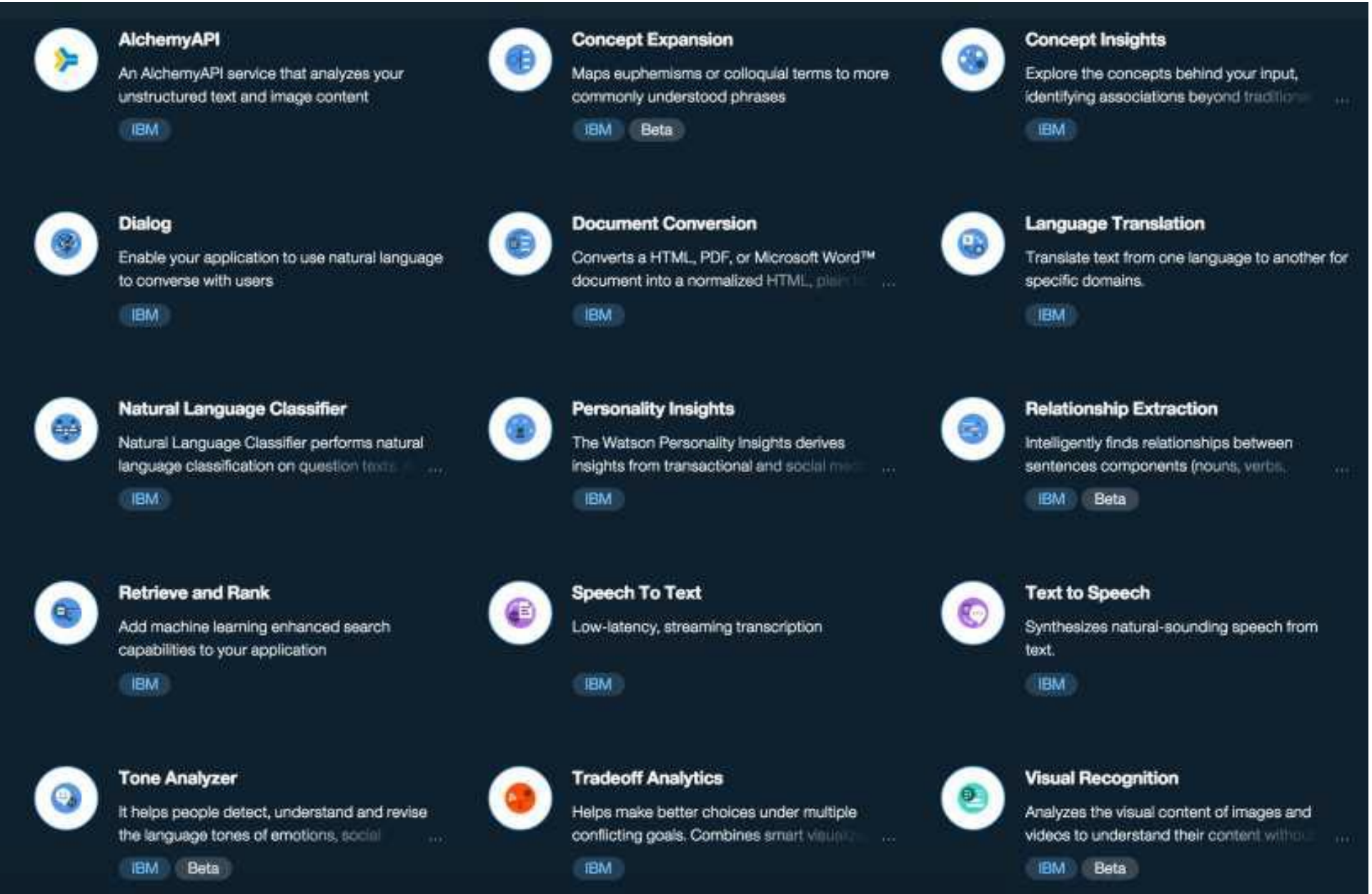
- Relatively cheap robot
- Lower quality
- Calling cloud cognitive services, e.g., face recognition
- Without a connection to the Internet is not working

Google Cloud

Google cloud provides APIs for computer vision, speech recognition, natural language processing, and translation.

- *Google Cloud Video Intelligence API* makes videos searchable and discoverable by extracting metadata, identifying key nouns, and annotating the content of the video.
- *Google Cloud Vision API* enables you to understand the content of an image including categories, objects and faces, words, and more. Face recognition is a common use of Vision API.
- *Google Cloud Speech API* enables you to convert audio to text by applying neural network models in an easy to use API.
- *Google Natural Language API* provides developers functionality to information about people, places, events and much more, mentioned in text documents, news articles or blog posts.
- *Google Cloud Translation API* lets developers convert text from a source language to a target language.

IBM Watson


AlchemyAPI
An AlchemyAPI service that analyzes your unstructured text and image content
IBM
Concept Expansion
Maps euphemisms or colloquial terms to more commonly understood phrases
IBM
Beta
Concept Insights
Explore the concepts behind your input, identifying associations beyond
IBM
Dialog
Enable your application to use natural language to converse with users
IBM
Document Conversion
Converts a HTML, PDF, or Microsoft Word™ document into a normalized HTML,
IBM
Language Translation
Translate text from one language to another for specific domains.
IBM
Natural Language Classifier
Natural Language Classifier performs natural language classification on question
IBM
Personality Insights
The Watson Personality Insights derives insights from transactional and social
IBM
Relationship Extraction
Intelligently finds relationships between sentences components (nouns, verbs,
IBM
Beta
Retrieve and Rank
Add machine learning enhanced search capabilities to your application
IBM
Speech To Text
Low-latency, streaming transcription
IBM
Text to Speech
Synthesizes natural-sounding speech from text.
IBM
Tone Analyzer
It helps people detect, understand and revise the language tones of emotions, social
IBM
Beta
Tradeoff Analytics
Helps make better choices under multiple conflicting goals. Combines smart visual
IBM
Visual Recognition
Analyzes the visual content of images and videos to understand their content
IBM
Beta

MicroSoft Azure

Cognitive Services APIs

Vision API
Computer Vision
Custom Vision Service
Face API
Forms Recognizer PREVIEW
Ink Recognizer PREVIEW
Video Indexer

Speech API
Speech Services
Speaker Recognition PREVIEW
Bing Speech API RETIRING
Translator Speech RETIRING

Search API
Bing News Search
Bing Video Search
Bing Web Search
Bing Autosuggest
Bing Custom Search
Bing Entity Search
Bing Image Search
Bing Visual Search
Bing Spell Check
Bing Local Business Search PREVIEW

Decision API
Anomaly Detector PREVIEW
Content Moderator
Personalizer PREVIEW

Language API
Language Understanding (LUIS)
QnA Maker
Text Analytics
Translator Text

OpenAI

MODEL	DESCRIPTION
GPT-4 Turbo and GPT-4	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
GPT-3.5 Turbo	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
DALL·E	A model that can generate and edit images given a natural language prompt
TTS	A set of models that can convert text into natural sounding spoken audio
Whisper	A model that can convert audio into text
Embeddings	A set of models that can convert text into a numerical form
Moderation	A fine-tuned model that can detect whether text may be sensitive or unsafe
GPT base	A set of models without instruction following that can understand as well as generate natural language or code
Deprecated	A full list of models that have been deprecated along with the suggested replacement

A dark side of the cloud services

- Cloud services could be very comfortable
- However, they are not:
  - because of their business model
  - each user must register
  - each call is charged
  - quality can be disputable, and the service rather freely collects data from users
- One can a get free period or free initial amount of calls
- Exception: Android platform can call Google cloud without any restrictions

Voice recognition from Android

https://github.com/andylucny/Recognize4PC